Complex Networks and Decentralized Search Algorithms

Jon Kleinberg*

Abstract. The study of complex networks has emerged over the past several years as a theme spanning many disciplines, ranging from mathematics and computer science to the social and biological sciences. A significant amount of recent work in this area has focused on the development of random graph models that capture some of the qualitative properties observed in large-scale network data; such models have the potential to help us reason, at a general level, about the ways in which real-world networks are organized.

We survey one particular line of network research, concerned with small-world phenomena and decentralized search algorithms, that illustrates this style of analysis. We begin by describing a well-known experiment that provided the first empirical basis for the "six degrees of separation" phenomenon in social networks; we then discuss some probabilistic network models motivated by this work, illustrating how these models lead to novel algorithmic and graph-theoretic questions, and how they are supported by recent empirical studies of large social networks.

Mathematics Subject Classification (2000). Primary 68R10; Secondary 05C80, 91D30.

Keywords. Random graphs, complex networks, search algorithms, social network analysis.

1. Introduction

Over the past decade, the study of complex networks has emerged as a theme running through research in a wide range of areas. The growth of the Internet and the World Wide Web has led computer scientists to seek ways to manage the complexity of these networks, and to help users navigate their vast information content. Social scientists have been confronted by social network data on a scale previously unimagined: datasets on communication within organizations, on collaboration in professional communities, and on relationships in financial domains. Biologists have delved into the interactions that define the pathways of a cell's metabolism, discovering that the network structure of these interactions can provide insight into fundamental biological processes. The drive to understand all

*Supported in part by a David and Lucile Packard Foundation Fellowship, a John D. and Catherine T. MacArthur Foundation Fellowship, and NSF grants CCF-0325453, IIS-0329064, CNS-0403340, and BCS-0537606.

these issues has resulted in what some have called a "new science of networks" — a phenomenological study of networks as they arise in the physical world, in the virtual world, and in society.

At a mathematical level, much of this work has been rooted in the study of *random graphs* [14], an area at the intersection of combinatorics and discrete probability that is concerned with the properties of graphs generated by random processes. While this has been an active topic of study since the work of Erdös and Rényi in the 1950s [26], the appearance of rich, large-scale network data in the 1990s stimulated a tremendous influx of researchers from many different communities. Much of this recent cross-disciplinary work has sought to develop random graph models that more tightly capture the qualitative properties found in large social, technological, and information networks; in many cases, these models are closely related to earlier work in the random graphs literature, but the issues arising in the motivating applications lead to new types of mathematical questions. For surveys covering different aspects of this general area, and in particular reflecting the various techniques of some of the different disciplines that have contributed to it, we refer the reader to recent review papers by Albert and Barabási [4], Bollobás [15], Kleinberg and Lawrence [39], Newman [52], and Strogatz [60], the volume of articles edited by Ben-Naim et al. [10], and the monographs by Dorogovtsev and Mendes [23] and Durrett [25], as well as books by Barabási [8] and Watts [62] aimed at more general audiences.

What does one hope to achieve from a probabilistic model of a complex network arising in the natural or social world? A basic strategy pursued in much of this research is to define a stylized network model, produced by a random mechanism that reflects the processes shaping the real network, and to show that this stylized model reproduces properties observed in the real network. Clearly the full range of factors that contribute to the observed structure will be too intricate to be fully captured by any simple model. But a finding based on a random graph formulation can help argue that the observed properties may have a simple underlying basis, even if their specifics are very complex. While it is crucial to realize the limitations of this type of activity — and not to read too much into the detailed conclusions drawn from a simple model — the development of such models has been a valuable means of proposing concrete, mathematically precise hypotheses about network structure and evolution that can then serve as starting points for further empirical investigation. And at its most effective, this process of modeling via random graphs can suggest novel types of qualitative network features — structures that people had not thought to define previously, and which become patterns to look for in new network datasets.

In the remainder of the present paper, we survey one line of work, motivated by the "small-world phenomenon" and some related search problems, that illustrates this style of analysis. We begin with a striking experiment by the social psychologist Stanley Milgram that frames the empirical issues very clearly [50, 61]; we describe a sequence of models based on random graphs that capture aspects of this phenomenon [64, 36, 37, 38, 63]; and we then discuss recent work that has identified some of the qualitative aspects of these models in large-scale network

data [1, 43, 49]. We conclude with some further extensions to these random graph models, discussing the results and questions that they lead to.

2. The Small-World Phenomenon

The small-world phenomenon – the principle that we are all linked by short chains of acquaintances, or "six degrees of separation" [29] – has long been the subject of anecdotal fascination among the general public, and more recently has become the subject of both experimental and theoretical research. At its most basic level, it is a statement about networks, and human social networks in particular; it concerns the graph with one node corresponding to each person in the world, and an edge joining two people if they know each other on a first-name basis. When we say that this graph is a "small world," we mean, informally, that almost every pair of nodes is connected by a path with an extremely small number of steps.

One could worry about whether this graph is precisely specified — for example, what exactly it means to know someone on a first-name basis — but however one fixes a working definition for this, it is clear that the resulting graph encodes an enormous amount of information about society in general. It is also clear that it would be essentially impossible to determine its structure precisely. How then could one hope to test, empirically, the claim that most pairs of nodes in this graph are connected by short paths?

The social psychologist Stanley Milgram [50, 61] took up this challenge in the 1960s, conducting an experiment to test the small-world property by having people explicitly construct paths through the social network defined by acquaintanceship. To this end, he chose a *target person* in the network, a stockbroker living in a suburb of Boston, and asked a collection of randomly chosen "starter" individuals each to forward a letter to the target. He provided the target's name, address, occupation, and some personal information, but stipulated that the participants could not mail the letter directly to the target; rather, each participant could only advance the letter by forwarding it to a single acquaintance that he or she knew on a first-name basis, with the goal of reaching the target as rapidly as possible. Each letter thus passed successively from one acquaintance to another, closing in on the stockbroker outside Boston.

The letters thus acted as virtual "tracers," mapping out paths through the social network. Milgram found that the median length among the completed paths was six, providing the first concrete evidence for the abundance of short paths connecting far-flung pairs of individuals in society, as well as supplying the basis for the number "six" in the resulting pop-cultural mantra. One needs to be careful in interpreting this finding, of course: many of the chains never reached the target, and the target himself was a relatively "high-status" individual who may have been easier to reach than an arbitrary person (see e.g. the recent critique by Kleinfeld [41]). But since Milgram's work, the overall conclusion has been accepted at least at a qualitative level: social networks tend to exhibit very short paths between essentially arbitrary pairs of nodes.

3. Basic Models of Small-World Networks

Why should social networks exhibit this type of a small-world property? Earlier we suggested that interesting empirical findings about networks often motivate the development of new random graph models, but we have to be careful in framing the issue here: a simple abundance of short paths is in fact something that most basic models of random graphs already "get right." As a paradigmatic example of such a result, consider the following theorem of Bollobás and de la Vega [14, 17].

Theorem 3.1 ([17]). *Fix a constant $k \geq 3$. If we choose uniformly at random from the set of all n-node graphs in which each node has degree exactly k, then with high probability every pair of nodes will be joined by a path of length $O(\log n)$.*

(Following standard notation and terminology, we say that the *degree* of a node is the number of edges incident to it. We say that a function is $O(f(n))$ if there is a constant c so that for all sufficiently large n, the function is bounded by $cf(n)$.) In fact, [17] states a much more detailed result concerning the dependence on n, but this will not be crucial for our purposes here.

Path lengths that are logarithmic in n — or more generally *polylogarithmic*, bounded by a polynomial function of $\log n$ — will be our "gold standard" in most of this discussion. We will keep the term "small world" itself informal; but we will consider a graph to be a small world, roughly, when all (or most) pairs of nodes are connected by paths of length polylogarithmic in n, since in such a case the path lengths are exponentially smaller than the number of nodes.

Watts and Strogatz [64] argued that there was something crucial missing from the picture provided by Theorem 3.1. A standard random graph (for example, as in Theorem 3.1) is locally very sparse; with reasonably high probability, none of the neighbors of a given node v are themselves neighbors of one another. But this is far from true in most naturally occurring networks: in real network data, many of a node's neighbors are joined to each other by edges. (For example, in a social network, many of our friends know each other.) Indeed, at an implicit level, this is a large part of what makes the small-world phenomenon surprising to many people when they first hear it: the social network appears from the local perspective of any one node to be highly "clustered," rather than the kind of branching tree-like structure that would more obviously reach many nodes along very short paths.

Thus, Watts and Strogatz proposed thinking about small-world networks as a kind of superposition: a structured, high-diameter network with a relatively small number of "random" links added in. As a model for social networks, the structured underlying network represents the "typical" social links that we form with the people who live near us, or who work with us; the additional random links are the chance, long-range connections that play a large role in creating short paths through the network as a whole.

This kind of hybrid random graph model had been studied earlier by Bollobás and Chung [16]; they showed that a small density of random links can indeed produce short paths very effectively. In particular they proved the following, among other results.

Theorem 3.2 ([16]). *Consider a graph G formed by adding a random matching to an n-node cycle. (In other words, we assume n is even, pair up the nodes on the cycle uniformly at random, and add edges between each of these node pairs.) With high probability, every pair of nodes will be joined by a path of length $O(\log n)$.*

Here too, Bollobás and Chung in fact proved a much more detailed bound on the path lengths; see [16] for further details.

This is quite close to the setting of the Watts-Strogatz work, who also considered cycles with random matchings as a model system for analysis. For our purposes here, we will begin with the following *grid-based model*, which is qualitatively very similar. We start with a two-dimensional $n \times n$ grid graph, and then for each node v, we add one extra directed edge to some other node w chosen uniformly at random. (We will refer to w as the *long-range contact* of v; to distinguish this, we will refer to the other neighbors of v, defined by the edges of the grid, as its *local contacts*.) Following the Watts-Strogatz framework, one can interpret this model as a metaphor for a social network embedded in an underlying physical space — people tend to know their geographic neighbors, as well as having friendships that span long distances. It is also closely related to *long-range percolation models*, though the questions we consider are fairly different; we discuss these connections in Section 7. For the present discussion, though, the essential feature of this model is its superposition of structured and random links, and it is important to note that the results to follow carry over directly to a wide range of variations on the model. Indeed, a significant part of what follows will be focused on a search for the most general framework in which to formulate these results.

4. Decentralized Search in Small-World Networks

Thus far we have been discussing purely structural issues; but if one thinks about it, the original Milgram experiment contains a striking algorithmic discovery as well: not only did short paths exist in the social network, but people, using knowledge only of their own acquaintances, were able to collectively construct paths to the target. This was a necessary consequence of the way Milgram formulated the task for his participants; if one really wanted the *shortest* path from a starting person to the target, one should have instructed the starter to forward a letter to *all* of his or her friends, who in turn should have forwarded the letter to all of their friends, and so forth. This "flooding" of the network would have reached the target as rapidly as possible; but for obvious reasons, such an experiment was not a feasible option. As a result, Milgram was forced to embark on the much more interesting experiment of constructing paths by "tunneling" through the network, with the letter advancing just one person at a time — a process that could well have failed to reach the target, even if a short path existed.

This algorithmic aspect of the small-world phenomenon raises fundamental questions — why should the social network have been structured so as to make this type of decentralized routing so effective? Clearly the network contained some type of "gradient" that helped participants guide messages toward the target, and

this is something that we can try to model; the goal would be to see whether decentralized routing can be proved to work in a simple random-graph model, and if so, to try extracting from this model some qualitative properties that distinguish networks in which this type of routing can succeed. It is worth noting that these issues reach far beyond the Milgram experiment or even social networks; routing with limited information is something that takes place in communication networks, in browsing behavior on the World Wide Web, in neurological networks, and in a number of other settings — so an understanding of the structural underpinnings of efficient decentralized routing is a question that spans all these domains.

To begin with, we need to be precise about what we mean by a decentralized algorithm. In the context of the grid-based model in the previous section, we will consider algorithms that seek to pass a message from a starting node s to a target node t, by advancing the message along edges. In each step of this process, the current message-holder v has knowledge of the underlying grid structure, the location of the target t on the grid, and its own long-range contact. The crucial point is that it does not know the long-range contacts of any other nodes. (Optionally, we can choose to have v know the path taken by the message thus far, but this will not be crucial in any of the results to follow.) Using this information, v must choose one of its network neighbors w to pass the message to; the process then continues from w. We will evaluate decentralized algorithms according to their *delivery time* — the expected number of steps required to reach the target, over a randomly generated set of long-range contacts, and randomly chosen starting and target nodes. Our goal will be to find algorithms with delivery times that are polylogarithmic in n.

It is interesting that while Watts and Strogatz proposed their model without the algorithmic aspect in mind, it is remarkably effective as a simple system in which to study the effectiveness of decentralized routing. Indeed, to be able to pose the question in a non-trivial way, one wants a network that is partially known to the algorithm and partially unknown — clearly in the Milgram experiment, as well as in other settings, individual nodes use knowledge not just of their own local connections, but also of certain global "reference frames" (comparable to the grid structure in our setting) in which the network is embedded. Furthermore, for the problem to be interesting, the "known" part of the network should be likely to contain no short path from the source to the target, but there should be a short path in the full network. The Watts-Strogatz model combines all these features in a minimal way, and thus allows us to consider how nodes can use what they know about the network structure to construct short paths.

Despite all this, the first result here is negative.

Theorem 4.1 ([36, 37]). *The delivery time of any decentralized algorithm in the grid-based model is* $\Omega(n^{2/3})$.

(We say that a function is $\Omega(f(n))$ if there is a constant c so that for infinitely many n, the function is at least $cf(n)$.)

This shows that there are simple models in which there can be an exponential separation between the lengths of paths and the delivery times of decentralized

algorithms to find these paths. However, it is clearly not the end of the story; rather, it says that the random links in the Watts-Strogatz model are somehow too "unstructured" to support the kind of decentralized routing that one found in the Milgram experiment. It also raises the question of finding a simple extension of the model in which efficient decentralized routing becomes possible.

To extend the model, we introduce one additional parameter $\alpha \geq 0$ that controls the extent to which the long-range links are correlated with the geometry of the underlying grid. First, for two nodes v and w, we define their *grid distance* $\rho(v, w)$ to be the number of edges in a shortest path between them on the grid. The idea behind the extended model is to have the long-range contacts favor nodes at smaller grid distance, where the bias is determined by α. Specifically, we define the *grid-based model with exponent* α as follows. We start with a two-dimensional $n \times n$ grid graph, and then for each node v, we add one extra directed edge to some other long-range contact; we choose w as the long-range contact for v with probability proportional to $\rho(v, w)^{-\alpha}$. Note that $\alpha = 0$ corresponds to the original Watts-Strogatz model, while large values of α produce networks in which essentially no edges span long distances on the grid.

We now have a continuum of models that can be studied, parameterized by α. When α is very small, the long-range links are "too random," and can't be used effectively by a decentralized algorithm; when α is large, the long-range links appear to be "not random enough," since they simply don't provide enough of the long-distance jumps that are needed to create a small world. Is there an optimal operating point for the network, where the distribution of long-range links is sufficiently balanced between these extremes to be of use to a decentralized routing algorithm?

In fact there is; as the following theorem shows, there is a unique value of α in the grid-based model for which a polylogarithmic delivery time is achievable.

Theorem 4.2 ([36, 37]). *(a) For $0 \leq \alpha < 2$, the delivery time of any decentralized algorithm in the grid-based model is $\Omega(n^{(2-\alpha)/3})$.*

(b) For $\alpha = 2$, there is a decentralized algorithm with delivery time $O(\log^2 n)$.

(c) For $\alpha > 2$, the delivery time of any decentralized algorithm in the grid-based model is $\Omega(n^{(\alpha-2)/(\alpha-1)})$.

(We note that the lower bounds in (a) and (c) hold even if each node has an arbitrary constant number of long-range contacts, rather than just one.)

The decentralized algorithm achieving the bound in (b) is very simple: each node simply forwards the message to a neighbor — long-range or local — whose grid distance to the target is as small as possible. (In other words, each node uses its long-range contact if this gets the message closer to the target on the grid; otherwise, it uses a local contact in the direction of the target.) The analysis of this algorithm proceeds by showing that, for a constant $\varepsilon > 0$, there is a probability of at least $\varepsilon / \log n$ in every step that the grid distance to the target will be halved. It is also worth noting that the proof can be directly adapted to a grid in any constant number of dimensions; an analogous trichotomy arises, with polylogarithmic delivery time achievable only when α is equal to the dimension.

At a more general level, the proof of Theorem 4.2(b) shows that the crucial property of exponent $\alpha = 2$ is the following: rather than producing long-range contacts that are uniformly distributed over the grid (as one gets from exponent $\alpha = 0$), it produces long-range contacts that are approximately uniformly distributed over "distance scales": the probability that the long-range contact of v is at a grid distance between 2^{j-1} and 2^j away from v is approximately the same for all values of j from 1 to $\frac{1}{2}\log n$.

From this property, one sees that there is a reasonable chance of halving the message's grid distance to the target, independent of how far away it currently is. The property also has an intuitively natural meaning in the context of the original Milgram experiment; subject to all the other simplifications made in the grid model, it says very roughly that decentralized routing can be effective when people have approximately the same density of acquaintances at many different levels of distance resolution. And finally, this approximate uniformity over distance scales is the type of qualitative property that we mentioned as a goal at the outset. It is something that we can search for in other models and in real network data — tasks that we undertake in the next two sections.

5. Decentralized Search in Other Models

Hierarchical Models. A natural variation on the model of the previous section is to suppose that the network is embedded in a hierarchy rather than a grid — in other words, that the nodes reside at the leaves of a complete b-ary tree, and the underlying "distance" between two nodes is based on the height of their lowest common ancestor in this tree.

There are a number of settings where such a model suggests itself. To begin with, follow-up work on the Milgram experiment found that most decisions made by participants on how to forward the letter were based on one of two kinds of cues: geographical and occupational [35]. And if a two-dimensional grid is natural as a simple abstraction for the role of geography, then a hierarchy is a reasonable, also simple, approximation of the way in which people categorize occupations. Another domain in which hierarchies arise naturally is in the relationships among Web pages: for example, a Web page about sequence analysis of the yeast genome could be classified as being about genetics, more generally about biology, and more generally still about science, while a Web page reviewing performances of Verdi's *Aida* could be classified as being about opera, more generally about music, and more generally still about the arts.

A natural assumption is that the density of links is lower for node pairs that are more widely separated in the underlying hierarchy, and this forms the basis for the following *hierarchical model with exponent* β. We begin with a complete b-ary tree having n leaves (and hence of height $h = \log_b n$). For two leaves v and w, let us define their *tree distance* $h(v, w)$ to be the height of their lowest common ancestor in the underlying tree. We now define the following random directed graph G on the set V of leaves: for a value k and for each node v in V, we construct k edges

out of v, choosing w as the endpoint of the i^{th} edge independently with probability proportional to $b^{-\beta h(v,w)}$. (We will refer to k as the *out-degree* of the model.)

Thus, β works much like α did in the grid-based model; when $\beta = 0$, we get uniform random selection, while larger values of β bias the selection more toward "nearby" nodes. Now, in this case, a decentralized search algorithm is given the locations of a starting node s and a target node t in the hierarchy, and it must construct a path from s to t, knowing only the edges out of nodes that it explicitly visits. Note that in defining the performance metric for a decentralized search algorithm in this model, we face a problem that we didn't encounter in the grid-based model: the graph G may not contain a path from s to t. Thus, we say that a decentralized algorithm here has delivery time $f(n)$ if, on a randomly generated n-node network, and with s and t chosen uniformly at random, the algorithm produces a path of length $O(f(n))$ with probability at least $1 - \varepsilon(n)$, where $\varepsilon(\cdot)$ is a function going to 0 as n increases.

We now have the following analogue of Theorem 4.2, establishing that there is a unique value of β for which polylogarithmic delivery time can be achieved when the network has polylogarithmic out-degree. This is achieved at $\beta = 1$, when the probability that v links to a node at tree distance h is almost uniform over choices of h. Also by analogy with the grid-based model, it suffices to use the simple "greedy" algorithm that always seeks to reduce the tree distance to the target by as much as possible.

Theorem 5.1 ([38]). *(a) In the hierarchical model with exponent $\beta = 1$ and out-degree $k = c \log^2 n$, for a sufficiently large constant c, there is a decentralized algorithm with polylogarithmic delivery time.*

(b) For every $\beta \neq 1$ and every polylogarithmic function $k(n)$, there is no decentralized algorithm in the hierarchical model with exponent β and out-degree $k(n)$ that achieves polylogarithmic delivery time.

Watts, Dodds, and Newman [63] independently proposed a model in which each node resides in several distinct hierarchies, reflecting the notion that participants in the small-world experiment were simultaneously taking into account several different notions of "proximity" to the target. Concretely, their model constructs a random graph G as follows. We begin with q distinct complete b-ary trees, for a constant q, and in each of these trees, we independently choose a random one-to-one mapping of the nodes onto the leaves. We then apply a version of the hierarchical model above, separately in each of the trees; the result is that each node of G acquires edges independently through its participation in each tree. (There are a few minor differences between their procedure within each hierarchy and the hierarchical model described above; in particular, they map multiple nodes to the same leaf in each hierarchy, and they generate each edge by choosing the tail uniformly at random, and then the head according to the hierarchical model. The result is that nodes will not in general all have the same degree.)

Precisely characterizing the power of decentralized search in this model, at an analytical level, is an open question, but Watts et al. describe a number of interesting findings obtained through simulation [63]. They study what is perhaps the

most natural search algorithm, in which the current message-holder forwards the message to its neighbor who is closest (in the sense of tree distance) to the target in any of the hierarchies. Using an empirical definition of efficiency on networks of several hundred thousand nodes, they examined the set of (β, q) pairs for which the search algorithm was efficient; they found that this "searchable region" was centered around values of $\beta \geq 1$ (but relatively close to 1), and on small constant values of q. (Setting q equal to 2 or 3 yielded the widest range of β for which efficient search was possible.) The resulting claim, at a qualitative level, is that efficient search is facilitated by having a small number of different ways to measure proximity of nodes, and by having a small bias toward nearby nodes in the construction of random edges.

Models based on Set Systems. One can imagine many other ways to construct networks in this general style — for example, placing nodes on both a hierarchy and a lattice simultaneously — and so it becomes natural to consider more general frameworks in which a range of these bounds on searchability might follow simultaneously from a single result. One such approach is based on constructing a random graph from an underlying set system, following the intuition that individuals in a social network often form connections because they are both members of the same small group [38]. In other words, two people might be more likely to form a link because they live in the same town, work in the same profession, have the same religious affiliation, or follow the work of the same obscure novelist.

Concretely, we start with a set of nodes V, and a collection of subsets $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ of V, which we will call the set of *groups*. It is hard to say much of interest for arbitrary set systems, but we would like our framework to include at least the collection of balls or subsquares in a grid, and the collection of rooted sub-trees in a hierarchy. Thus we consider set systems that satisfy some simple combinatorial properties shared by these two types of collections. Specifically, for constants $\lambda < 1$ and $\kappa > 1$, we impose the following three properties.

(i) The full set V is one of the groups.

(ii) If S_i is a group of size $g \geq 2$ containing a node v, then there is a group $S_j \subseteq S_i$ containing v that is strictly smaller than S_i, but has size at least $\min(\lambda g, g - 1)$.

(iii) If $S_{i_1}, S_{i_2}, S_{i_3}, \ldots$ are groups that all have size at most g and all contain a common node v, then their union has size at most κg.

The most interesting property here is (iii), which can viewed as a type of "bounded growth" requirement; one can easily verify that it (along with (i) and (ii)) holds for the set of balls in a grid and the set of rooted sub-trees in a hierarchy.

Given a collection of groups, we construct a random graph as follows. For nodes v and w, we define $g(v, w)$ to be the size of the smallest group containing both of them — this will serve as a notion of "distance" between v and w. For a fixed exponent γ and out-degree value k, we construct k edges out of each node v, choosing w as the endpoint of the i^{th} edge from v independently with probability

proportional to $g(v,w)^{-\gamma}$. We will refer to this as the *group-based model* with set system $\mathcal{S}$, exponent γ, and out-degree k. A decentralized search algorithm in such a random graph is given knowledge of the full set system, and the identity of a target node; but it only learns the links out of a node v when it reaches v. We now have the following theorem.

Theorem 5.2 ([38]). *(a) Given an arbitrary set system $\mathcal{S}$ satisfying properties (i), (ii), and (iii), there is a decentralized algorithm with polylogarithmic delivery time in the group-based model with set system $\mathcal{S}$, exponent $\gamma = 1$, and out-degree $k = c \log^2 n$, for a sufficiently large constant c.*

(b) For every set system $\mathcal{S}$ satisfying properties (i), (ii), and (iii), every $\gamma < 1$, and every polylogarithmic function $k(n)$, there is no decentralized algorithm achieving polylogarithmic delivery time in the group-based model with set system $\mathcal{S}$, exponent γ and out-degree $k(n)$.

In other words, efficient decentralized search is possible when nodes link to each other with probability inversely proportional to the size of the smallest group containing both of them. As a simple concrete example, if the groups are the balls in a two-dimensional grid, then the size of the smallest group containing two nodes at distance ρ is proportional to ρ^2, and so the link probability indicated by Theorem 5.2(a) is proportional to ρ^{-2}; this yields an analogue of Theorem 4.2(b), the inverse-square result for grids. (The present setting is not exactly the same as the one there; here, we do not automatically include the edges of the original grid when constructing the graph, but we construct a larger number of edges out of each node.)

Simple examples show that one cannot directly formulate a general negative result in this model for the case of exponents $\gamma > 1$ [38]. At a higher level, the group-based model is clearly not the only way to generalize the results thus far; in the next section we will discuss one other recent approach, and the development of other general models is a natural direction for further research.

6. Design Principles and Network Data

In addition to their formulation as basic questions about search algorithms in graphs, the models we have been discussing thus far have been used as design principles in file-sharing systems; and they have been found to capture some of the large-scale structure of human social networks as reflected in on-line data.

Peer-to-Peer Systems and Focused Web Crawling. A recurring theme in recent work on complex networks is the way in which simple probabilistic models can rapidly become design principles for new types of networked systems. In the case of small-world networks, one observes this phenomenon in the development of protocols for peer-to-peer file sharing. The design of such protocols has become an active topic of research in the area of computer systems, motivated in part by the explosion of popular interest in peer-to-peer applications following the emergence

of Napster and music file-sharing in 1999. The goal of such applications was to allow a large collection of users to share the content residing on their personal computers, and in their initial conception, the systems supporting them were based on a centralized index that simply stored, in a single place, the files that all users possessed. This way, queries for a particular piece of content could be checked against this index, and routed to the computer containing the appropriate file.

The music-sharing application of these systems, of course, ran into significant legal difficulties; but independent of the economic and intellectual property issues raised by this particular application, it is clear that systems allowing large user communities to share content have a much broader range of potential, less controversial uses, provided they can be structured in a robust and efficient way. This has stimulated much subsequent study in the research community, focusing on *decentralized* approaches in which one seeks file-sharing solutions that do not rely on a single centralized index of all the content.

In this decentralized version of the problem, the crux of the challenge is clear: each user has certain files on his or her own computer, but there is no single place that contains a global list of all these files; if someone poses a query looking for a specific piece of content, how can we efficiently determine which user (if any) possesses a copy of it? Without a central index, we are in a setting very much like that of the Milgram experiment: users must pose the query to a subset of their immediate network neighbors, who in turn can forward the query to some of their neighbors, and so forth. And this is where small-world models have played a role: a number of approaches to this problem have tried to explicitly set up the network on which the protocol operates so that its structure makes efficient decentralized search possible. We refer the reader to the surveys by Aspnes and Shah [6] and Lua et al. [44] for general reviews of this body of work, and the work of Clarke et al. (as described in [32]), Zhang et al. [67], Malkhi et al. [45], and Manku et al. [46] for more specific discussions of the relationship to small-world networks.

A related set of issues comes up in the design of *focused Web crawlers*. Whereas standard Web search engines first compile an enormous index of Web pages, and then answer queries by referring to this index, a focused crawler attempts to locate pages on a specific topic by following hyperlinks from one page to another, without first compiling an index. Again, the underlying issue here is the design of decentralized search algorithms, in this case for the setting of the Web: when searching for relevant pages without global knowledge of the network, what are the most effective rules for for deciding which links to follow? Motivated by these issues, Menczer [49] studied the extent to which the hierarchical model described in the previous section captures the patterns of linkage in large-scale Web data, using the hierarchical organization of topics provided by the Open Directory.

Social Network Data. The previous two applications — peer-to-peer systems and focused Web crawling — are both concerned with the structure of computer and information networks, although in both cases there are obvious social forces underlying their construction. Recent work has also investigated the extent to which the models described in the previous sections are actually reflected in data

on human social networks. In other words, these small-world models make very concrete claims about the ways in which networks should be organized to support efficient search, but it is not *a priori* clear whether or not real networks are organized in such ways. Two recent studies of this flavor have both focused on social networks that exist in on-line environments — as with the previous applications, we again see an intertwining of social and technological networks, but in these cases the emphasis is on the social component, with the on-line aspect mainly providing an opportune means of performing fine-grained analysis on a large scale.

In one study of this flavor, Adamic and Adar [1] considered the e-mail network of a corporate research lab: they collected data over a period of time, and defined an edge between two people who exchanged at least a certain number of messages during this period. They overlaid the resulting network on a set system representing the organizational structure, with a set for each subgroup of the lab's organizational hierarchy. Among other findings, they showed that the probability of a link between individuals v and w scaled approximately proportional to $g(v,w)^{-3/4}$, compared with the value $g(v,w)^{-1}$ for efficient search from Theorem 5.2(a). (As above, $g(v,w)$ denotes the size of the smallest group containing both v and w.) Thus, interactions in their data spanned large groups at a slightly higher frequency than the optimum for decentralized search. Of course, the e-mail network was not explicitly designed to support decentralized search, although one can speculate about whether there were implicit factors shaping the network into a structure that was easy to search; in any case, it is interesting that the behavior of the links with respect to the collection of groups is approximately aligned with the form predicted by the earlier theorems.

An even closer correlation with the structure predicted for efficient search was found in a large-scale study by Liben-Nowell et al. [43]. They considered LiveJournal, a highly active on-line community with several million participants, in which members communicate with one another, update personal on-line diaries, and post messages to community discussions. LiveJournal is a particularly appealing domain for studying the geographic distribution of links, because members provide explicit links to their friends in the system, and a large subset (roughly half a million at the time of the study in [43]) also provide a hometown in the continental U.S. As a result, one has the opportunity to investigate, over a very large population, how the density of social network links decays with distance.

A non-trivial technical challenge that must be overcome in order to relate this data to the earlier models is that the population density of the U.S. is extremely non-uniform, and this makes it difficult to interpret predictions based on a model in which nodes are distributed uniformly over a grid. The generalization to group structures in the previous section is one way to handle non-uniformity; Liben-Nowell et al. propose an alternative generalization, *rank-based friendships*, that they argue may be more suitable to the geographic data here [43]. In the rank-based friendship model, one has a set of n people assigned to locations on a two-dimensional grid, where each grid node may have an arbitrary positive number of people assigned to it. By analogy with the grid-based model from Section 4, each person v chooses a *local contact* arbitrarily in each of the four neighboring grid

nodes, and then chooses an additional *long-range contact* as follows. First, v ranks all other people in order of their distance to herself (breaking ties in some canonical way); we let $\text{rank}_v(w)$ denote the position of w in v's ordered list, and say that w is at *rank* r with respect to v. v then chooses w as her long-range contact with probability proportional to $1/\text{rank}_v(w)$.

Note that this model generalizes the grid-based model of Section 4, in the sense that the grid-based model with the inverse-square distribution corresponds to rank-based friendship in which there is one person resident at each grid node. However, the rank-based friendship construction is well-defined for any population density, and Liben-Nowell et al. prove that it supports efficient decentralized search in general. They analyze a decentralized greedy algorithm that always forwards the message to a grid node as close as possible to the target's; and they define the *delivery time* in this case to be the expected number of steps needed to reach the grid node containing the target. (So we can imagine that the task here is to route the message to the hometown of the target, rather than the target himself; this is also consistent with the data available from LiveJournal, which only localizes people to the level of towns.)

Theorem 6.1 ([43]). *For an arbitrary poopulation density on a grid, the expected delivery time of the decentralized greedy algorithm in the rank-based friendship model is* $O(\log^3 n)$.

On the LiveJournal data. Liben-Nowell et al. examine the fraction of friendships (v, w) where w is at rank r with respect to v. They find that this fraction is very close to inverse linear in r, in close alignment with the predictions of the rank-based friendship model.

This finding is notable for several reasons. First, as with the e-mail network considered by Adamic and Adar, there is no *a priori* reason to believe that a large, apparently amorphous social network should correspond so closely to a distribution predicted by a simple model for efficient decentralized search. Second, geography is playing a strong role here despite the fact that LiveJournal is an on-line system in which there are no explicit limitations on forming links with people arbitrarily far away; as a result, one might have (incorrectly) conjectured that it would be difficult to detect the traces of geographic proximity in such data. And more generally, the analytical results of this section and the previous ones have been based on highly stylized models that nonetheless make very specific predictions about the theoretical "optimum" for search; to see these concrete predictions approximately borne out on real social network data is striking, and it suggests that there may be deeper phenomena yet to be discovered here.

7. Further Results on Small-World Networks and Decentralized Search

Long-Range Percolation. The grid-based models we have been considering are closely related to the problem of *long-range percolation*. In the basic version

of long-range percolation, one takes the infinite d-dimensional integer lattice $\mathbf{Z}^d$, and for each pair of nodes (v, w) one includes an undirected edge between them independently with probability $\rho(v, w)^{-\alpha}$, where $\rho(v, w)$ is the grid distance between v and w and $\alpha \geq 0$ is a parameter of the model. Note that there are some small differences from the grid-based model described in Section 4: the graph is infinite, it is undirected, its nodes do not all have the same degree, and it does not automatically include edges between nearest neighbors on the lattice. In addition to these, a broader difference is in the nature of the questions investigated, with the initial work on long-range percolation focusing on the range of parameters for which an infinite connected component was likely to exist [3, 51, 57].

Motivated in part by the interest in small-world networks, work on long-range percolation began to investigate diameter issues — the maximum D for which every node is connected by a path of at most D steps. Benjamini and Berger [11] studied this problem in one dimension, modifying the model so that the graph is finite (restricted to the integers $\{1, 2, \ldots, n\}$), and so that edges are guaranteed to exist between adjacent integers. (They also studied the case in which the distance $\rho(\cdot, \cdot)$ is defined by assuming that the integers are "wrapped" into a cycle, so that $\rho(i, j)$ is not $|j - i|$ but $\min(|j - i|, n - |j - i|)$.) Their work was followed by results of Coppersmith et al. [20] and Biskup [13], who obtained sharper bounds in some cases and considered higher-dimensional lattices as well, in which the node set is $\{1, 2, \ldots, n\}^d$. As a result of this work, we know that the diameter of the graph changes qualitatively at the "critical values" $\alpha = d$ and $\alpha = 2d$. In particular, with high probability, the diameter is constant when $\alpha < d$ (due in essence to a result of [12]), is proportional to $\log n / \log \log n$ when $\alpha = d$ [20], is polylogarithmic in n when $d < \alpha < 2d$ (with an essentially tight bound provided in [13]), and is lower-bounded by a polynomial in n when $\alpha > 2d$ [11, 20]. The case $\alpha = 2d$ is largely open, and conjectured to have diameter polynomial in n with high probability [11, 13]. It is also open whether the diameter for $\alpha > 2d$ is in fact linear in n; this has been proved for the one-dimensional case [11] and conjectured to hold for higher dimensions as well [11, 13, 20].

This pair of transitions at $\alpha = d$ and $\alpha = 2d$ was observed in a somewhat different setting by Kempe et al. [34], resolving a conjecture of Demers et al. [21] on the behavior of *gossip algorithms*. In this model, there are nodes located on the finite d-dimensional lattice $\{1, 2, \ldots, n\}^d$, and in each time step each node v picks a single other node and tells everything it currently knows to w; node w is selected as the recipient of this information with probability proportional to $\rho(v, w)^{-\alpha}$. Information originating at one node thus spreads to other nodes, relayed in an epidemic fashion over time. Now, if a single node v initially possesses a new piece of information at time 0, how long will it take before knowledge of this information has spread to a given node w? The main result of [34] is that the time required for this is polylogarithmic in n for $\alpha \leq d$, is polylogarithmic in $\rho(v, w)$ but independent of n for $d < \alpha < 2d$, and is polynomial in $\rho(v, w)$ for $\alpha > 2d$. Here too the case $\alpha = 2d$ is not well understood, which is interesting because this transitional value has particular importance in applications of gossip algorithms to distributed computing systems [54]. (See [34] for partial results concerning

$\alpha = 2d$.)

For the specific grid-based model described in Section 4, Martel and Nguyen showed that with high probability the diameter is proportional to $\log n$ for $\alpha \leq d$, in the d-dimensional case [48]. They also identified transitions at $\alpha = d$ and $\alpha = 2d$ analogous to the case of long-range percolation [53]. In particular, their results show that while decentralized search can construct a path of length $O(\log^2 n)$ when $\alpha = d$, there in fact exist paths that are shorter by a logarithmic factor. (Note also the contrast with the corresponding results for the long-range percolation model when $\alpha \leq d$; in the grid-based model, the out-degree of each node is bounded by a constant, so a diameter proportional to $\log n$ is the smallest one could hope for; in the case of long-range percolation, on the other hand, the node degrees will be unbounded, allowing for smaller diameters.)

Decentralized Search with Additional Information. A number of papers have studied the power of decentralized search algorithms that are provided with small amounts of additional information [28, 42, 47, 48, 66]. Whereas the model of decentralized algorithms in Section 4 charged unit cost to the algorithm for each node visited, the models in these subsequent papers make the following distinction: a node may "consult" a small number of nearby nodes, and then based on what it learns from this consultation, it chooses a node to forward the messages to. In bounding the number of steps taken by the algorithm, only the message-forwarding operations are counted, not the consultation.

In particular, Lebhar and Schabanel [42] consider an algorithm in which the node currently holding the message consults a set S of up to $O(\log n)$ nodes within a small number of steps of it; after this, it forwards the message along a path to the node w in S that is closest to the target in grid distance. They show that, in total, the expected number of nodes consulted by this process is $O(\log^2 n)$ (as in the decentralized algorithm from Section 4), and that the actual path constructed to the target has only $O(\log n(\log\log n)^2)$ steps.

Manku, Naor, and Wieder [47] consider a simpler algorithm in the long-range percolation model on the d-dimensional lattice $\{1, 2, \ldots, n\}^d$ with $\alpha = d$. Note that nodes here will have unbounded degrees — proportional to $\log n$ in expectation, rather than constant as in the grid-based model. Manku et al. analyze a *neighbor-of-neighbor* search algorithm in which the current message-holder v consults each of its neighbors to learn the set S of all of *their* neighbors; v then forwards the message along the two-step path to the node in S that lies closest to the target. They show that with high probability, this algorithm produces a path to the target of at most $O(\log n/\log\log n)$ steps, matching the bound of Coppersmith et al. [20] on the diameter of this network. Moreover, they show that the basic greedy algorithm, which simply forwards the message to the neighbor closest to the target, requires an expected number of steps proportional to $\log n$ to reach the target. Thus, one step of lookahead provides an asymptotic improvement in delivery time; and since one step of lookahead yields path lengths matching the diameter, additional lookahead does not offer any further asymptotic improvements.

Thus, the results of Manku et al. provide a rather sharp characterization of the

power of lookahead in the long-range percolation model at the exponent $\alpha = d$ that allows for efficient decentralized search; determining a similarly precise delineation on the power of lookahead in the grid-based model (extending the aforementioned results of Lebhar and Schabanel) is an interesting open question.

Small-World Networks Built on Arbitrary Underlying Graphs. The results in Section 5 describe various methods for constructing searchable networks based on underlying structures other than d-dimensional grids. In several recent papers, a number of further structures have been proposed as "scaffolds" for small-world networks [9, 27, 31, 53, 59].

In principle, one can consider adding long-range edges to any underlying graph G; Fraigniaud [27] asks whether any G can be converted through such a process into a network that is efficiently searchable by a greedy algorithm. Specifically, suppose we choose a distribution over long-range contacts for each node of G, and we use this to generate a random graph G' by adding a single long-range edge out of each node of G. We then consider the natural greedy algorithm for forwarding the message to a target t: the current message-holder passes the message to a neighbor that has the shortest path to the target as measured in G (not in G'). Is it the case that for every graph G, there is a distribution over long-range contacts such that this algorithm has a delivery time that is polylogarithmic in n?

This question is open in general; note that the challenge in resolving it comes from the fact that a single choice of distribution per node must work (in expectation) over any possible destination, and that even if the graph G' has nicely-structured short paths, the search algorithm is constrained to behave "greedily" in the original graph G. Fraigniaud answers the question in the affirmative for graphs of bounded tree-width as well as graphs in which there is no induced cycle of greater than a fixed length [27]; he also discusses some respects in which such underlying graphs are qualitatively consistent with observed properties of social networks. Duchon et al. answer the question in the affirmative for graphs satisfying a certain "bounded growth rate" property [24].

Slivkins [59] considers a different setting, in which nodes are embedded in an underlying metric space. He shows that if the metric is *doubling*, in the sense that every ball can be covered by a constant number of balls of half the radius (see e.g. [7, 30]), then there is a model such that each node generates a polylogarithmic number of long-range contacts from specified distributions, and a decentralized algorithm is then able to achieve a polylogarithmic delivery time. (Some of the logarithmic dependence here is on the *aspect ratio* of the metric — the ratio of the largest to the smallest distance — but it is possible to avoid this dependence in the bound on the delivery time. See [59] for further details on this issue.)

Finally, other work has studied search algorithms that exploit differences in node degrees. There are indications that people navigating social structures, in settings such as small-world experiments, take into account the fact that certain of their acquaintances simply know a large number of people [22]. Similarly, in peer-to-peer networks, it is also the case that certain nodes have an unusually large number of neighbors, and may thus be more useful in helping to forward queries.

Adamic et al. [2] formalize these considerations by studying a random graph model in which high-degree nodes are relatively abundant, and decentralized search algorithms only have access to information about degrees of neighboring nodes, not to any embedding of the graph (spatial or otherwise). Through simulation, they find that for certain models, knowledge of degrees provides an improvement in search performance.

Simsek and Jensen [58] consider a model which combines spatial embedding with variable node degrees. Specifically, they study a variant of the grid-based model from Section 4 in which nodes have widely varying degrees, and a decentralized algorithm has access both to the locations of its neighbors and to their degrees. Through simulation, they find that a heuristic taking both these factors into account can perform more efficiently than decentralized algorithms using only one of these sources of information. Finding the optimal way to combine location and degree information in decentralized search, and understanding the range of networks that are searchable under such optimal strategies, is an interesting direction for further research.

8. Conclusion

We have followed a particular strand of research running through the topic of complex networks, concerned with short paths and the ability of decentralized algorithms to find them. As suggested initially, the sequence of ideas here is characteristic of the flavor of research in this area: an experiment in the social sciences that highlights a fundamental and non-obvious property of networks (efficient searchability, in this case); a sequence of random graph models and accompanying analysis that seeks to capture this notion in a simple and stylized form; a set of measurements on large-scale network data that parallels the properties of the models, in some cases to a surprising extent; and a range of connections to further results and questions in algorithms, graph theory, and discrete probability.

To indicate some of the further directions in which research on this topic could proceed, we conclude with a list of open questions and issues related to small-world networks and decentralized search. Some of these questions have already come up implicitly in the discussion thus far, so one goal of this list is to collect a number of these questions in a single place. Other questions here, however, bring in issues that reach beyond the context of the earlier sections. And as with any list of open questions, we must mention a few caveats: the questions here take different forms, since some are concretely specified while other are more designed to suggest problems in need of a precise formulation; the questions are not independent, in that the answer to one might well suggest ways of approaching others; and several of the questions may well become more interesting if the underlying model or formulation is slightly varied or tweaked.

1. Variation in Node Degrees. As we discussed at the end of the previous section, decentralized search in models that combine wide variation in node degrees

with some kind of spatial embedding is an interesting issue that is not well understood. Simsek and Jensen's study [58] of this issue left open the question of proving bounds on the efficiency of decentralized algorithms. For example, consider the d-dimensional grid-based model with exponent α, and suppose that rather than constructing a fixed number of long-range contacts for each node, we draw the number of long-range contacts for each node v independently from a given probability distribution. To be concrete, we could consider a distribution in which one selects k long-range contacts with probability proportional to $k^{-\delta}$ for a constant δ.

We now have a family of grid-based models parameterized by α and δ, and we can study the performance of decentralized search algorithms that know not only the long-range contacts out of the current node, but also the degrees of the neighboring nodes. Decentralized selection of a neighbor for forwarding the message has a stochastic optimization aspect here, balancing the goal of forwarding to a node to the target with the goal of forwarding to a high-degree node. We can now ask the general question of how the delivery time of decentralized algorithms varies in both α and δ. Note that it is quite possible this question becomes more interesting if we vary the model so that long-range links are undirected; this way, a node with a large degree is both easy to find and also very useful once it is found. (In a directed version, a node with large out-degree may be relatively useless simply because it has low in-degree and so is unlikely to be found.)

2. The case of $\alpha = 2d$. In both the grid-based model and the related long-range percolation models, very little is known about the diameter of the graph when α is equal to twice the dimension. (It appears that a similar question arises in other versions of the group-based models from Section 5, when nodes form links with probability inversely proportional to the square of the size of the smallest group containing both of them.) Resolving the behavior of the diameter would shed light on this transitional point, which lies at the juncture between "small worlds" and "large worlds." This open question also manifests itself in the gossip problem discussed in Section 7, where we noted that the transitional value $\alpha = 2d$ arises in distributed computing applications (see the discussion in [34, 54]).

3, Paths of Logarithmic Length. It would be interesting to know whether there is a decentralized algorithm in the d-dimensional grid-based model, at the "searchable exponent" $\alpha = d$, that could construct paths of length $O(\log n)$ while visiting only a polylogarithmic number of nodes. This would improve the result of Lebhar and Schabanel [42] to an asymptotically tight bound on path length.

4. Small-World Networks with an Arbitrary Base Graph. It would also be interesting to resolve the open problem of Fraigniaud [27] described in Section 7, formalizing the question of whether any graph can be turned into an efficiently searchable small world by appropriately adding long-range links.

5. Extending the Group-Based Model. Theorem 5.2 on the group-based model contained a positive result generalizing the ones for grids and hierarchies, and it contained a general negative result for the case when long-range connection were "too long-range" (i.e. with exponent $\gamma < 1$). However, it does not fully generalize the results for grids and hierarchies, because there are set systems satisfying conditions (i), (ii), and (iii) of the theorem for which efficient decentralized search is possible even for exponents $\gamma > 1$. It would be interesting to find a variation on these three properties that still generalizes grids and hierarchies in a natural way, and for which $\gamma = 1$ is the unique exponent at which efficient decentralized search is possible.

6. Multiple Hierarchies. Obtaining provable bounds for decentralized search in the "multiple hierarchies" model of Watts, Dodds, and Newman [63] is also an open question. Such results could form an interesting parallel with the findings they discovered through simulation. With some small modifications to the model of Watts et al., one can cast it in the group-based model of Section 5, and so it is entirely possible that progress on this question and the previous could be closely connected.

7. The Evolution of Searchable Networks. The remaining questions have a more general flavor, where much of the challenge is the formalization of the underlying issue. To begin with, the current models supporting efficient decentralized search are essentially *static*, in that they describe how the underlying network is organized without suggesting how it might have evolved into this state. What kinds of growth processes or selective pressures might exist to cause networks to become more efficiently searchable? Interesting network evolution models addressing this question have been proposed by Clauset and Moore [19] and by Sandberg [56], both based on feedback mechanisms by which nodes repeatedly perform decentralized searches and in the process partially "rewire" the network. Obtaining provable guarantees for these models, or variations on them, is an open question. A number of peer-to-peer file-sharing systems include similar feedback mechanisms, achieving good performance in practice. Freenet [18] is a good example of such a system, and the relationship of its feedback mechanism to the evolution of small-world networks is studied by Zhang et al. [67].

Game theory may provide another promising set of techniques for studying the evolution of small-world networks. A growing body of recent work has considered game-theoretic models of network formation, in which agents controlling nodes and edges interact strategically to construct a graph — the basic question is to understand what types of structures emerge when each agent is motivated by self-interest. For surveys of this area, see [5, 33, 65]. In the present case, it would be interesting to understand whether there are ways to define incentives such that the collective outcome of self-interested behavior would be a searchable small-world network.

8. Decentralized Search in the Presence of Incentives. Game-theoretic notions can also provide insight not just into the growth of a network, but also into the processes that operate on it. A topic of interest in the peer-to-peer community, as well as in the design of on-line communities, is the way in which the incentives offered to the members of the system influence the extent to which they are willing to forward queries and information. In the case of decentralized search, suppose that there is some utility associated with routing the message from the starting node to the target, and intermediate nodes behave strategically, demanding compensation for their participation in the construction of the path. How do results on decentralized path formation change when such behavior is incorporated into the model?

In [40], this question is made precise in a setting where the underlying network is a random tree, constructed via a branching process. It would be interesting to consider analogous issues in richer classes of networks.

9. Reconstruction. The networks we have considered here have all been embedded in some underlying "reference frame" — grids, hierarchies, or set systems — and most of our analysis has been predicated on a model in which the network is presented together with this embedding. This makes sense in many contexts; recall, for example, the discussion from Section 6 of network data explicitly embedded in Web topic directories [49], corporate hierarchies [1], or the geography of the U.S. [43]. In some cases, however, we may be presented with just the network itself, and the goal is to determine whether it has a natural embedding into a spatial or hierarchical structure, and to recover this embedding if it exists. For example, we may have data on communication within an organization, and the goal is to reconstruct the hierarchical structure under the assumption that the frequency of communication decreases according to a hierarchical model — or to reconstruct the positions of the nodes under the assumption that the frequency of communication decreases with distance according to a grid-based or rank-based model.

One can formulate many specific questions of this flavor. For example, given a network known to be generated by the grid-based model with a given exponent α, can we approximately reconstruct the positions of the nodes on the grid? What if we are not told the exponent? Can we determine whether a given network was more likely to have been generated from a grid-based model with exponent α or α'? Or what if there are multiple long-range contacts per node, and we are only shown the long-range edges, not the local edges? A parallel set of questions can be asked for the hierarchical model.

Questions of this type have been considered by Sandberg [55], who reports on the results of computational experiments but leaves open the problem of obtaining provable guarantees. Benjamini and Berger [11] pose related questions, including the problem of reconstructing the dimension d of the underlying lattice when presented with a graph generated by long-range percolation on a finite piece of $\mathbf{Z}^d$.

10. Comparing Network Datasets. As we saw earlier, the models proposed in Sections 4 and 5 suggest a general perspective from which to analyze network datasets, by studying the way in which the density of links decays with increasing distance or increasing group size (e.g. [1, 43]). One could naturally use this style of analysis to compare related network datasets — for example taking the patterns of communication within k different organizations (as Adamic and Adar did for the corporate lab they studied), and determining exponents $\gamma_1, \gamma_2, \ldots, \gamma_k$ for each such that the probability of a link between individuals v and w in a group of size g scales approximately as $g^{-\gamma_i}$ in the i^{th} organization. Differences among these exponents would suggest structural differences between the organizations at a global level — communication in some is more long-range, while in others it is more clustered at the low levels of the hierarchy. It would be interesting to understand whether these differences in turn were naturally reflected in other aspects of the organizations' behavior and performance.

More generally, large-scale social, technological, and information networks are sufficiently complex objects that the guiding principles provided by simple models seem crucial for our understanding of them. The perspective suggested here has offered one such collection of principles, highlighting in particular the ways in which these networks are intertwined with the spatial and organizational structures that they inhabit. One can hope that as we gather an increasing range of different perspectives, our understanding of complex networks will continue to deepen into a rich and informative theory.

References

[1] L. Adamic, E. Adar. How to search a social network. Social Networks, 27(3):187-203, July 2005.

[2] L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman. Search in Power-Law Networks. Phys. Rev. E, 64 46135 (2001).

[3] M. Aizenman, C.M. Newman. Discontinuity of the Percolation Density in One-Dimensional $1/|x-y|^2$ Percolation Models. Commun. Math. Phys. 107(1986).

[4] R. Albert, A.-L. Barabási. Statistical mechanics of complex networks. Review of Modern Physics 74, 47-97 (2002)

[5] E. Anshelevich. Network Design and Management with Strategic Agents. Ph.D. thesis, Cornell University, 2005.

[6] J. Aspnes, G. Shah. Distributed data structures for P2P systems. in Theoretical and Algorithmic Aspects of Sensor, Ad Hoc Wireless and Peer-to-Peer Networks (Jie Wu, ed.), CRC Press, 2005.

[7] P. Assouad. Plongements lipschitziens dans $\mathbf{R}^n$. Bull. Soc. Math. France 111(1983).

[8] A.-L. Barabási. Linked. Perseus 2002.

[9] L. Barrière, P. Fraigniaud, E. Kranakis, D. Krizanc. Efficient Routing in Networks with Long Range Contacts. Proceedings of DISC 2001.

[10] Eli Ben-Naim, Hans Frauenfelder, Zoltan Toroczkai, eds. Complex Networks Springer Lecture Notes in Physics (vol. 650), 2004.

[11] I. Benjamini, N. Berger. The diameter of long-range percolation clusters on finite cycles. Random Structures and Algorithms 19(2001).

[12] I. Benjamini, H. Kesten, Y. Peres, O. Schramm. Geometry of the uniform spanning forest: transitions in dimensions $4, 8, 12, \ldots$. Annals of Mathematics 2(2004).

[13] M. Biskup. On the scaling of the chemical distance in long range percolation models. Ann. Probab. 32 (2004)

[14] B. Bollobás. Random Graphs (2nd edition). Cambridge University Press, 2001.

[15] B. Bollobás. Mathematical results on scale-free random graphs. In Handbook of Graphs and Networks (Stefan Bornholdt, Hans Georg Schuster, eds.), Wiley 2004.

[16] B. Bollobás, F.R.K. Chung. The diameter of a cycle plus a random matching. SIAM J. Discrete Math. 1(1988).

[17] B. Bollobás, W. F. de la Vega. The diameter of random regular graphs. Combinatorica 2(1982) 125-134

[18] I. Clarke, O. Sandberg, B. Wiley, T. Hong. Freenet: A Distributed Anonymous Information Storage and Retrieval System. International Workshop on Design Issues in Anonymity and Unobservability, 2000.

[19] A. Clauset, C. Moore. How Do Networks Become Navigable? preprint at arxiv.org, 2003.

[20] D. Coppersmith, D. Gamarnik, M. Sviridenko. The diameter of a long-range percolation graph. Random Structures and Algorithms 21(2002).

[21] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. ACM SIGOPS Operating Systems Review, vol. 22, no. 1, January 1988.

[22] P. Dodds, R. Muhamad, D. J. Watts. An Experimental Study of Search in Global Social Networks. Science 301(2003), 827.

[23] S.N. Dorogovtsev, J.F.F. Mendes. Evolution of Networks: from biological networks to the Internet and WWW. Oxford University Press, 2003.

[24] P. Duchon, N. Hanusse, E. Lebhar, N. Schabanel. Could any graph be turned into a small world? Theoretical Computer Science, to appear.

[25] R. Durrett. Random Graph Dynamics. Cambridge University Press, 2006.

[26] P. Erdös and A. Rényi. On the Evolution of Random Graphs. Mat. Kutato Int. Kozl 5 (1960), 17-60.

[27] P. Fraigniaud. A New Perspective on the Small-World Phenomenon: Greedy Routing in Tree-Decomposed Graphs Proc. 13th Annual European Symposium on Algorithms (ESA), 2005.

[28] P. Fraigniaud, C. Gavoille, and C. Paul. Eclecticism shrinks even small worlds. Proceedings of 23rd Annual Symposium on Principles of Distributed Computing, 2004.

[29] J. Guare. Six Degrees of Separation: A Play. Vintage Books, 1990.

[30] Anupam Gupta, Robert Krauthgamer and James R. Lee Bounded geometries, fractals, and low-distortion embeddings. Proc. 44th IEEE Symposium on Foundations of Computer Science, 2003.

[31] D. Higham. Greedy Pathlengths and Small World Graphs. University of Strathclyde Mathematics Research Report 08(2002).

[32] T. Hong. Performance. In Peer-to-Peer: Harnessing the Power of Disruptive Technologies (A. Oram, editor), O'Reilly and Associates, 2001.

[33] M. Jackson. A Survey of Models of Network Formation: Stability and Efficiency. In Group Formation in Economics; Networks, Clubs and Coalitions, (G. Demange and M. Wooders, eds.), Cambridge University Press, 2004.

[34] D. Kempe, J. Kleinberg, A. Demers. Spatial Gossip and Resource Location Protocols. Proc. 33rd ACM Symposium on Theory of Computing, 2001.

[35] P. Killworth and H. Bernard. Reverse small world experiment. Social Networks 1(1978).

[36] J. Kleinberg. Navigation in a Small World. Nature 406(2000).

[37] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Proc. 32nd ACM Symposium on Theory of Computing, 2000.

[38] J. Kleinberg. Small-World Phenomena and the Dynamics of Information. Advances in Neural Information Processing Systems (NIPS) 14, 2001.

[39] J. Kleinberg, S. Lawrence. The Structure of the Web. Science 294(2001).

[40] J. Kleinberg, P. Raghavan. Query Incentive Networks. Proc. 46th IEEE Symposium on Foundations of Computer Science, 2005.

[41] J. Kleinfeld. Could it be a Big World After All? The 'Six Degrees of Separation' Myth. Society, April 2002.

[42] E. Lebhar, N. Schabanel, Almost optimal decentralized routing in long-range contact networks, Proceedings of the International Colloquium on Automata, Languages, and Programming (ICALP), 2004.

[43] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins. Geographic routing in social networks. Proc. Natl. Acad. Sci. USA, 102(Aug 2005).

[44] E-K Lua, J. Crowcroft, M. Pias, R. Sharma and S. Lim. A Survey and Comparison of Peer-to-Peer Overlay Network Schemes, *IEEE Communications Surveys and Tutorials*, 7(2005).

[45] D. Malkhi, M. Naor, D. Ratajczak. Viceroy: a scalable and dynamic emulation of the butterfly. Proceedings of 21st Annual Symposium on Principles of Distributed Computing, 2002.

[46] G. S. Manku, M. Bawa, P. Raghavan. Symphony: Distributed hashing in a small world. Proc. 4th USENIX Symposium on Internet Technologies and Systems, 2003.

[47] G. Manku, M. Naor, and U. Wieder. Know Thy Neighbor's Neighbor: The Power of Lookahead in Randomized P2P Networks. Proc. of ACM Symp. on Theory of Computing (STOC), 2004.

[48] C. Martel, V. Nguyen. Analyzing Kleinberg's (and other) small-world models. Proceedings of 23rd ACM Symposium on Principles of Distributed Computing, 2004.

[49] F. Menczer. Growing and Navigating the Small World Web by Local Content. Proc. Natl. Acad. Sci. USA 99(22): 14014-14019, 2002

[50] S. Milgram, The small world problem. Psychology Today 1(1967).

[51] C.M. Newman, L.S. Schulman. One Dimensional $1/|j-i|^s$ Percolation Models: The Existence of a Transition for $s \leq 2$. Commun. Math. Phys. 104(1986)

[52] M.E.J. Newman. The structure and function of complex networks. SIAM Review, 45:167–256, 2003.

[53] V. Nguyen and C. Martel. Analyzing and characterizing small-world graphs. Proceedings of ACM-SIAM symposium on Discrete Algorithms, 2005.

[54] R. van Renesse, K. P. Birman, W. Vogels. Astrolabe: A robust and scalable technology for distributed system monitoring, management, and data mining. ACM Trans. Computer Sys. 21(2003).

[55] O. Sandberg. Distributed Routing in Small-World Networks. Algorithm Engineering and Experiments (ALENEX), 2006.

[56] O. Sandberg. Searching a Small World. Licentiate thesis, Chalmers University, 2005.

[57] L.S. Schulman. Long-range percolation in one dimension. J. Phys. A 16, no. 17, 1986.

[58] O. Simsek and D. Jensen. Decentralized search in networks using homophily and degree disparity. Proc. 19th International Joint Conference on Artificial Intelligence, 2005.

[59] A. Slivkins. Distance Estimation and Object Location via Rings of Neighbors. Proceedings of 24th Annual Symposium on Principles of Distributed Computing, 2005.

[60] S. Strogatz. Exploring complex networks. Nature 410(2001), 268.

[61] J. Travers and S. Milgram. An experimental study of the small world problem. Sociometry 32(1969).

[62] Duncan J. Watts. Six Degrees: The Science of a Connected Age, W. W. Norton, 2003.

[63] D. J. Watts, P. S. Dodds, M. E. J. Newman. Identity and Search in Social Networks. Science, 296, 1302-1305, 2002.

[64] Watts, D. J. and S. H. Strogatz. Collective dynamics of 'small-world' networks. Nature 393(1998).

[65] T. Wexler. Pricing Games with Selfish Users. Ph.D. thesis, Cornell University, 2005.

[66] J. Zeng, W.-J. Hsu, J. Wang. Near Optimal Routing in a Small-World Network with Augmented Local Awareness Parallel and Distributed Processing and Applications: Third International Symposium (ISPA), 2005.

[67] H. Zhang, A. Goel, R. Govindan. Using the Small-World Model to Improve Freenet Performance. Proc. IEEE Infocom, 2002.

Jon Kleinberg
Department of Computer Science
Cornell University
Ithaca NY 14853 USA